Symbiotic Interfaces For Wearable Face Recognition

Bradley A. Singletary and Thad E. Starner
College Of Computing, Georgia Institute of Technology, Atlanta, GA 30332
{bas,thad}@cc.gatech.edu

Abstract

We introduce a wearable face detection method that exploits constraints in face scale and orientation imposed by the proximity of participants in near social interactions. Using this method we describe a wearable system that perceives "social engagement," i.e., when the wearer begins to interact with other individuals. One possible application is improving the interfaces of portable consumer electronics, such as cellular phones, to avoid interrupting the user during face-to-face interactions.

Our experimental system proved $> 90\%$ accurate when tested on wearable video data captured at a professional conference. Over three hundred individuals were captured, and the data was separated into independent training and test sets.

A goal is to incorporate user interface in mobile machine recognition systems to improve performance. The user may provide real-time feedback to the system or may subtly cue the system through typical daily activities, such as turning to face a speaker, as to when conditions for recognition are favorable.

1 Introduction

In casual social interaction, it is easy to forget the names and identities of those we meet. The consequences can range from the need to be reintroduced to the "opportunity cost" of a missed business contact. At organized social gatherings, such as professional conferences, name tags are used to assist attendees' memories. Recently, electronic name tags have been used to transfer, index, and remember contact information for attendees [Borovoy et al., 1996]. For everyday situations where convention-style name tags are inappropriate, a wearable face recognition system may provide face-name associations and aid in recall of prior interactions with the person standing in front of the wearable user [Farringdon and Oni, 2000, Starner et al., 1997, Brzezowski et al., 1996, Iordanoglou et al., 2000]. Currently, such systems are computationally complex and create a drain on the limited battery resources of a wearable computer. However, when a conversant is socially engaged with the user, a weak constraint may be exploited for face recognition. Specifically, search over scale and orientation may be limited to that typical of the near social interaction distances. Thus, we desire a lightweight system that can detect social engagement and indicate that face recognition is appropriate.

Wearable computers must balance their interfaces against human burden. For example, if the wearable computer interrupts its user during a social interaction (e.g. to alert him to a wireless telephone call), the conversation may be disrupted by the intrusion. Detection of social engagement allows for blocking or delaying interruptions appropriately during a conversation.

The above applications motivate our work in attempting to recognize social engagement. To visually identify social engagement, we wish to use features endemic of that social process. Eye fixation, patterns of change in head orientation, social conversational distance, and change in visual spatial content may be relevant [Selker et al., 2001, Reeves, 1993, Hall, 1963]. For now, as we are uncertain which features are appropriate for recognition, we induce a set of behaviors to assist the computer. Specifically, the wearer aligns x's on an head-up display with the eyes of the subject to be recognized. As we learn more about the applicability of our method from our sample data set, we will extend our recognition algorithms to include non-induced behaviors.

While there are many face detection, localization, and recognition algorithms in the literature that were considered as potential solutions to our problem [Feraud et al., 2001, Rowley et al., 1998, Schneiderman and Kanade, 2000, Sung and Poggio, 1998, Leung et al., 1995], our task is to recognize social engagement in context of human behavior and the environment. Face presence may be one of the most important features, but it is not the only feature useful for segmenting engagement. In examination of 10 standard face databases ($> 19,000$ images), we found that background contents had little variation. By comparison, scenes obtained from a body-worn camera in everyday life contained highly varied scene backgrounds. In addition to the presence of the face, we would like to exploit the movement of the face with respect to the wearer's camera. Given prior work on the visual modeling of human interaction [Oliver et al., 1998, Ivanov et al., 1999, Moore, 2000, Starner and Pentland, 1998, Starner et al., 1998, Nefian, 1999], we chose hidden Markov Models(HMMs) as the basis of our recognition system.

2 Engagement Dataset

We collected video data from a wearable camera at an academic conference, a setting representative of social interaction of the wearer and new acquaintances. The capture environment was highly unconstrained and ranged from direct sunlight to darkened conference hall. Approximately 300 subjects were captured one or more times over 10 hours. The images in Figure 1 are locations in the video annotated by the wearer to be faces. Our prototype wearable camera video capture system (see Figure 2) consists of: a color camera, an infrared(IR) sensitive black and white camera, a low-power IR illuminator, two digital video(DV) recorder decks, one video character generator, one audio tone generator, a Sony Glasstron head-up display, and four lithium ion camcorder batteries. The output of one camera is split with a low-power video distribution amplifier and displayed in one eye of the head mount display. The signal is annotated with two 'x' characters spaced and centered horizontally then placed one third of the way from the top of the video frames (Figure 2). The other copy of the signal is saved to DV tape. To capture face data, the wearer of the vest approaches a subject and aligns the person's eyes with the two 'x' characters. The 'x' characters represent known locations for a subject's eyes to appear in the video feed. The marks and lens focus are ideally calibrated to be appropriate for footage taken at normal conversational distances from the subject. Once the marks are aligned, the wearer pushes a button that injects an easily detected tone into the DV deck's audio channel for later recovery. The audio tones serve as ground-truth markers for training purposes.

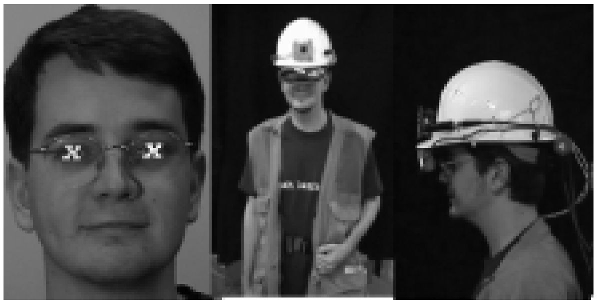

Figure 2: Marks for user alignment and face capture apparatus

3 Method

The video data was automatically extracted into 2 second partitions and divided into two classes using frames annotated by the wearer. The two classes were "engagement" and "other". Due to the fact that the wearer annotation was highly incomplete, we had to filter frontal facial interactions from the other class by hand. This editing was also used to exclude non-participants from the experiment. As may be expected, the number of engagement gestures per hour of interaction was much smaller than the number of examples in the garbage class.

Since the wearer lined up two x's with the eyes of a viewed subject, the presence of a face could safely be guaranteed to be framed by a 360x360 subregion of the 720x480 DV frame at the annotated locations in the video. Faces present at engagement were large with respect to the subregion. We first convert to greyscale, deinterlace, and correct non-squareness of the image pixels in the subregion. We downsampled the preprocessed region of video to 22x22 images using the linear heat equations to gaussian diffuse each level of the pyramid before subsampling to the next level. Each resulting frame/element in a 2-second gesture example is one 22x22 greyscale subregion (484 element vector). We model the face class by a 3 state Left-Right HMM as shown in Figure 3. The other class was much more complex to model and required a 6 state ergodic model to capture the interplay of garbage types of scenes as shown in Figure 3. We plot the mean values of the state output probabilities. The presence of a face seems important for acceptance by the face model. The first state contains a rough face-like blob and is followed by a confused state that likely represents the alignment portion of our gesture. The final state is

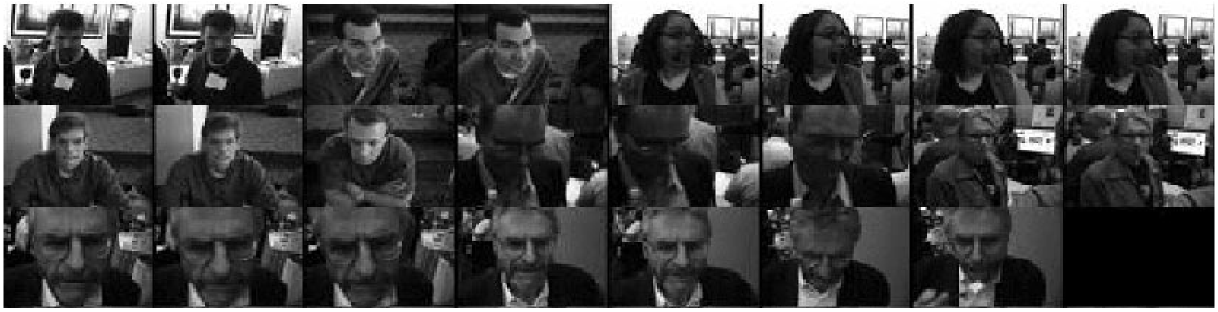

Figure 1: Representative data set

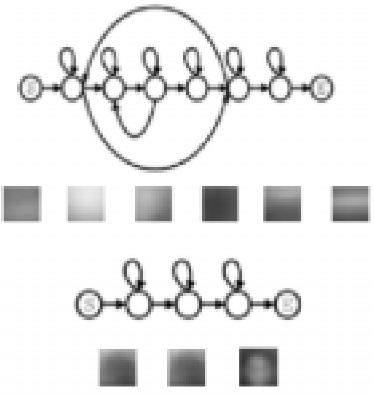

Figure 3: Other and Engagement classes

Table 1: Accuracy and confusion for engagement detection

experiment	*training set*	*independent test*
22x22 video stream	89.71%	90.10%

train confusion, N=843	*engagement*	*other*
engagement	82.1%(128)	17.9%(28)
other	8.6%(63)	91.3%(665)

test confusion, N=411	*engagement*	*other*
engagement	83.3%(50)	16.7%(10)
other	8.7%(30)	91.3%(314)

clearly face-like, with much sharper features than the first state and would be consistent with conversational engagement. Looking at the other class model, we see images that look like horizons and very dark or light scenes. The complexity of the model allowed wider variations in scene without loss in accuracy. Finally, it is important to note that modeling background could be improved by building location specific models as features specific to an environment could be better represented [Rungsarityotin and Starner, 2000].

4 Results and Evaluation Metrics

Accuracy results and confusion matrices are shown in Table 1. In wearable computing, battery life and processor speed are at a premium, resulting in a very specific evaluation criteria for our system. How effective is leveraging detection of social engagement as compared to continuously running face recognition? If we were to construct a wearable face recognition system using our engagement detector, we would combine the social engagement detector with a scale-tuned localizer and a face recognizer. The cost of the social engagement detector must be sufficiently small to allow for the larger costs of localization and recognition. This is described by the inequality

$$z - R_a * a \geq R_b * b$$

where $z := 1$ is the total resources available, a is the fixed cost of running engagement detection once in sec/frames, b is the fixed cost of running localization and recognition methods once in sec/frames, and R_a and R_b are the rate at which we can supply the respective detectors with frames in frames/sec, respectively. However, R_b has a maximum value determined by either the fraction of false positives U_{fp} multiplied by the maximum input frame rate or the rate at which the user wants to be advised of the identity of a conversant R_{ui}. Thus,

$$R_b * b \geq max\{R_a * U_{fp}, R_{ui}\} * b$$

Note that fixating the camera on a true face could cause up to R_a frames per second to be delivered to the face recognizer. However, we assume that the user does not want to be updated this quickly or repeatedly (i.e. $R_{ui} << R_a$). We also assume that our rate of false positives will almost always be greater than the rate the user wants to be informed, leaving us with

$$1 - R_a * a \geq R_a * U_{fp} * b$$

For comparison purposes, we will assume that the average time per frame of processing for the localization and recognition process can be represented by some multiple of the average detection time (i.e. $b = c * a$). Thus, for a given multiplier c, we can determine the maximum rate of false positives allowable by the face detection process.

$$U_{fp} \leq \frac{1}{R_a * a * c} - \frac{1}{c}$$

Note that if $c \leq 1$, then the localization and recognition process runs faster than the face detection process. This situation would imply that performing face detection separately from face localization and recognition would not save processing time (i.e. localization and recognition should run continually - again, if real-time face recognition is the primary goal).

Given a false positive rate U_{fp}, we can solve the equation to determine the maximum allowable time for the localization and recognition process as compared to the detection process.

$$c \leq \frac{1}{R_a * a * U_{fp}} - \frac{1}{U_{fp}}$$

Thus, we have a set of heuristics for determining when the separation of face detection and face localization and recognition is profitable.

5 Discussion and Applications

Applying the metric from the previous section to our experimental results, we let $U_{fp} = .13$, $R_a = 30$, $a = \frac{1}{60}$ and solving for c we get $c \leq 7.69$. Thus any recognition method used may be up to 7.69 times slower than the engagement detection method and will have a limiting frame rate of about four frames per second. Given that our detection algorithm runs at 30fps, and our knowledge that principal component analysis based face recognition and alignment can run faster than roughly four times a second, we feel that engagement detection can be a successful foundation for wearable face recognition. Post-filtering outputs of detection may help eliminate false positives before recognition [Feraud et al., 2001]. Due to the face-like appearance of the final state of the HMM, it is likely that the output of our method could provide a reasonable first estimate of location to fine grain localization.

We are working on modeling other modalities of engagement behavior. Mounting further sensors on the user may be useful for improving engagement detection. For example, Selker [Selker et al., 2001] proposes an eye fixation detector. It may be the case that eye fixation is indicative of social engagement. Two parties meeting for the first time will usually look to see whom they are meeting. Another modality we think useful is sound. For instance, personal utterances like “hello, my name is ...” are common to social engagement. A simple forward-looking range-sensor like sonar might help in disambiguating range. Also, a vision based walking/not-walking classifier was constructed, with favorable results in detection of walking, but has not yet been integrated. Detection of head stillness and other interest indicators will likely reduce false positives.[Reeves, 1993]

6 Conclusion

We described a platform built to capture video from a wearable user’s perspective and detailed a method for efficient engagement detection. We tested our system in a representative scenario and devised a metric for evaluating it’s efficacy as part of a face recognition scheme. In doing so, we demonstrated how the design of user interfaces that are aware of social contexts and constraints can positively affect recognition systems on the body. Finally, we have described how the detection of social engagement may be used, in its own right, to improve interfaces on portable consumer devices.

References

[Borovoy et al., 1996] Borovoy, R., McDonald, M., Martin, F., and Resnick, M. (1996). Things that blink: A computationally augmented name tag. *IBM Systems Journal*, 35(3).

[Brzezowski et al., 1996] Brzezowski, S., Dunn, C. M., and Vetter, M. (1996). Integrated portable system for suspect identification and tracking. In DePersia, A. T., Yeager, S., and Ortiz, S., editors, *SPIE:Surveillance and Assessment Technologies for Law Enforcement.*

[Farringdon and Oni, 2000] Farringdon, J. and Oni, V. (2000). Visually augmented memory. In *Fourth International Symposium on Wearable Computers*, Atlanta, GA. IEEE.

[Feraud et al., 2001] Feraud, R., Bernier, O. J., Viallet, J.-E., and Collobert, M. (2001). A fast and accurate face detector based on neural networks. *Pattern Analysis and Machine Intelligence*, 23(1):42–53.

[Hall, 1963] Hall, E. T. (1963). *The Silent Language*. Doubleday.

[Harrison et al., 1994] Harrison, B. L., Ishii, H., and Chignell, M. (1994). An empirical study on orientation of shared workspaces and interpersonal spaces in video-mediated collaboration. Technical Report OTP-94-2, University of Toronto, Ontario Telepresence Project.

[Iordanoglou et al., 2000] Iordanoglou, C., Jonsson, K., Kittler, J., and Matas, J. (2000). Wearable face recognition aid. In *Interntional Conference on Acoustics, Speech, and Signal Processing*. IEEE.

[Ivanov et al., 1999] Ivanov, Y., Stauffer, C., Bobic, A., and Grimson, E. (1999). Video surveillance of interactions. In *CVPR Workshop on Visual Surveillance*, Fort Collins, CO. IEEE.

[Leung et al., 1995] Leung, T. K., Burl, M. C., and Perona, P. (1995). Finding faces in cluttered scenes using random labelled graph matching. In *5th Inter. Conference on Computer Vision*.

[Moore, 2000] Moore, D. J. (2000). *Vision-based recognition of actions using context*. PhD thesis, Georgia Institute of Technology, Atlanta, GA.

[Nefian, 1999] Nefian, A. (1999). *A hidden Markov model-based approach for face detection and recognition*. PhD thesis, Georgia Institute of Technology, Atlanta, GA.

[Oliver et al., 1998] Oliver, N., Rosario, B., and Pentland, A. (1998). Statistical modeling of human interactions. In *CVPR Workshop on Interpretation of Visual Motion*, pages 39–46, Santa Barbara, CA. IEEE.

[Reeves, 1993] Reeves, J. (1993). The face of interest. *Motivation and Emotion*, 17(4).

[Rowley et al., 1998] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1).

[Rungsarityotin and Starner, 2000] Rungsarityotin, W. and Starner, T. (2000). Finding location using omnidirectional video on a wearable computing platform. In *International Symposium on Wearable Computing*, Atlanta, GA. IEEE.

[Schneiderman and Kanade, 2000] Schneiderman, H. and Kanade, T. (2000). A statistical model for 3d object detection applied to faces and cars. In *Computer Vision and Pattern Recognition*. IEEE.

[Selker et al., 2001] Selker, T., Lockerd, A., and Martinez, J. (2001). Eye-r, a glasses-mounted eye motion detection interface. In *to appear CHI2001*. ACM.

[Starner et al., 1997] Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R. W., and Pentland, A. (1997). Augmented reality through wearable computing. *Presence special issue on Augmented Reality*.

[Starner and Pentland, 1998] Starner, T. and Pentland, A. (1998). Real-time American sign language recognition using desktop and wearable computer based video. *Pattern Analysis and Machine Intelligence*.

[Starner et al., 1998] Starner, T., Schiele, B., and Pentland, A. (1998). Visual contextual awareness in wearable computing. In *International Symposium on Wearable Computing*.

[Sung and Poggio, 1998] Sung, K. K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence*, 20(1):39–51.